Remote mining: from clustering to DTM

J. García-Gutiérrez[1], F. Martínez-Álvarez[2], D. Laguna-Ruiz[1] and J. C. Riquelme[1]

[1]Department of Computer Science, University of Seville, Spain.
{jgarcia, dlaguna, riquelme}@lsi.us.es

[2]Area of Computer Science, Pablo Olavide University, Spain.
fmaralv@upo.es

Abstract

LIDAR data acquisition is becoming an indispensable task for terrain characterization in large surfaces. In Mediterranean woods this job results hard due to the great variety of heights and forms, as well as sparse vegetation that they present. A new data mining-based approach is proposed with the aim of classifying LIDAR data clouds as a first step in DTM generation. The developed methodology consists in a multi-step iterative process that splits the data into different classes (ground and low/med/high vegetation) by means of a clustering algorithm. This method has been tested on three different areas of the southern Spain with successful results, verging on 80% hits.

Keywords: LIDAR, DTM, clustering.

1. Introduction

The Regional Ministry for the Environment of Andalusia, the regional government in the South of Spain, owns two LIDAR sensors. Recently, this public entity has decided to develop a software system to improve the cartographic and environmental services using this technological advance because of the LIDAR's well-known capacity to create 3D models with high quality. In this context it is necessary to be able to obtain a DTM from LIDAR data point clouds and to distinguish ground and non-ground (vegetation) impact. Nowadays, most of the software used to perform this kind of work is based on proprietary systems like Terrascan *(Terrasolid Limited 2002)*. Our goal has been to develop free software in the near future to classify LIDAR data. As an initial step our study has been centered on applying data mining techniques (k-means clustering) to a LIDAR point cloud in order to obtain a digital elevation model (DEM).

Others authors have worked in how to build DTM from LIDAR data previously. A method for filtering laser data (Vosselman 2000) is proposed closely related to the erosion operator used for mathematical grey scale morphology. This method is based on height differences in a representative training dataset, then filter functions are derived that either preserve important terrain characteristics or minimize the number of classification errors. The work (Haugeraud and Harding 2001) propose a method for deforestation to identify ground and non-ground points based on the geometry of surface in the neighborhood of each return. Zhang (Zhang *et al.* 2003) uses a progressive method to erase non-ground points based on a threshold to study height differences among points.

In more recent times, the paper (Sithole and Vosselman 2005) makes classifications, by using a segmentation process in the LIDAR data point cloud. Then, every segment is classified on the basis of the geometric relations to the rest of segments. Bartels (Bartels *et al* 2006) presents a new filter based on statistical moments from data cloud to distinguish ground and non-ground points in an efficient way. In addition, in (Evans and Hudak 2007), an iterative multiscale algorithm for classifying LIDAR returns that exceed positive surface curvature thresholds have been developed. The authors maintain that the results show very few commission errors and high quality models.

Most of the approaches to develop DTM are based on grid techniques or on some kind of preliminary process like rasterization. This kind of techniques usually introduces distortion in the system that is studied and it can be an error source. In this situation, data mining techniques can be applied for optimal results because:

- It does not need any preprocess that could produce errors in the results.
- It can be easily applied to big datasets as, for example, the LIDAR data clouds.

Data mining is defined by Piatetski-Shapiro (Piatetski-Shapiro *et al.* 1991) as: *"the set of techniques that are concerned with finding patterns in data which are interesting (according to some user-defined measure of interestingness, e.g., with coverage above the requested threshold) and valid (according to some user defined measure of validity, e.g., classification accuracy)".* These kinds of techniques can be applied to any data source in general and it can particularly be applied on LIDAR data without loss of precision.

Clustering is one a widely used data mining techniques. There are lots of clustering strategies, but perhaps the most extended method is k-means. It has been applied on bioinformatics, pattern recognition and even in remote sensing and LIDAR data, too. In this way, we can find approaches like Filin's (Filin 2004) which makes DSM by studying the angles among neighbors. k-means classical algorithm is used by Filin's approach applied to the attribute set obtained from a data cloud. In this way, every point has a set of angles with its neighbors as a result of Delaunay's triangularization. From this data and the position of every point, the author builds clusters to identify each surface. Other approaches use k-means as help to segment data and get the individual trees. Thus, Morsdorf (Morsdorf *et al.* 2004) chooses local maxima inside the cloud as the initial point of every cluster. Then the algorithm builds clusters surrounding each maximum and with this, it gets the vegetation structure, an important parameter in fire risk assessment and fire behaviour modelling. Others techniques from data mining like neural networks are used with a similar purpose. Thus, Fernandes (Fernandes et al. 2005) developed a one-layer perceptron to classify signals from terrestrial LIDAR automatically in order to discover forest fire in early phases. The authors maintain their approach has detection efficiency of 93% and a false alarm percentage of 0.041%.

This paper is concerned with the separation of ground and vegetation points in Mediterranean woods from LIDAR data. The main novelty of this approach is the applying of clustering techniques to the build of DTM and concretely the build of a digital elevation model (DEM) from a previous deep classification.

The paper is organised as follows: In Section 2, the zone under study is presented and our approach is detailed. Section 3 presents results on high, medium and low resolution LIDAR data. The paper discusses results, proposes future avenues and concludes in Section 4.

2. Method

Our method is based on a multi-step k-means clustering applied to a LIDAR data cloud. Each step divides the original data cloud in two sets. Each set identifies the points for a possible classification: ground, low vegetation, med vegetation and high vegetation. We lean on the silhouette function to decide if it is possible to keep on dividing data. In the next paragraphs we describe the k-means algorithm, the silhoutte function and a deeper description of our approach is shown.

The *k*-means algorithm was originally presented by MacQueen (1968). For each cluster, its centroid is used as the most representative point, where the centroid μ_j of a group of elements x_j is defined as the centre of gravity of all the elements comprising the cluster.

The aim is to minimize intra-cluster variance, or the squared error function:

$$V = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left| x_j - \mu_i \right|^2 \tag{1}$$

k-means is the most popular method to perform clustering. It is an efficient and scalable method especially useful to deal with large datasets. It presents a computational complexity *O(nkt)*, where *n* is the number of objects, *k* the number of clusters and *t*, the number of iterations. A local optimum is reached when *k*<<*n* and *t*<< *n*, which is a very common situation.

The selection of an optimum number of clusters is still an open task. Recently, several approaches have been developed in order to determine this number (Hamerly and Elkan, 2003; Yan and Ye, 2007) but its application has been demonstrated to be useful only in individual areas. In this sense, the silhouette function (Kaufmann and Rouseeuw, 1990) provides a measure of the cluster's separation and can be used as a general-purpose method.

Let's consider an item *i* (already clustered) that belongs to the cluster A. We evaluate the average dissimilarity of *i* to all the other objects of A is evaluated and denoted by *a(i).* Analogously, the average dissimilarity of *i* to all the objects of *B* is called *dis(i, B).* The next step consists of computing *dis(i, B)* for every $B \neq A$ and, subsequently, the smallest dissimilarity is chosen and noted by *b(i)*=min{*dis(i, B)*}, $B \neq A$. Thus, *b(i)* represents the dissimilarity of *i* to its neighbour cluster. Finally, to determine how well a point is clustered, the silhouette function, shown in equation (3), is applied:

$$silh(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2}$$

Its value varies between –1 and +1, where +1 denotes clear cluster separation and –1 marks points with questionable cluster assignment. If cluster *A* is a set containing a single member, then *silh(i)* is not defined and the most neutral choice is to set *silh(i) = 0.* The objective function is the average of *silh(i)* over the *N* objects to be classified, and the best clustering is reached when the above mentioned function is maximized.

2. 1. Algorithm

The classification method is based on three iterations to classify all data. The first iteration classifies high vegetation. If the resulted clustering has a good silhouette value, it continues with medium vegetation and finally low vegetation. The unclassified data at last iteration is the ground data. In Figure 1, the process description can be seen for i-th iteration. Thus, iteration is divided into four steps. The first step takes the raw LIDAR data as input and builds a matrix as output with the minimum height of the terrain in every cell. To obtain good results it is necessary to set the cell size as a parameter for the algorithm. This parameter will determinate the size of the terrain contained in a cell of the matrix. This will be very important in the next step.

Once the matrix has been built, we go into step 2. For each point from the raw LIDAR cloud, a new measure is added, the biggest height difference that is calculated between the point height and the neighbour cells in the matrix, where the cell that could contain the point is included as a neighbour too. This new measure needs to know which cells are neighbours. A new parameter ε has to be given. This parameter can be defined as the biggest distance between the cell that includes the point and a possible neighbour cell. This parameter is related closely to the resolution of the data cloud and it will define the portion of terrain that is processed, together with the parameter in the paragraph before, to calculate the maximum height difference among points. It is important to realize parameter ε will be bigger for LIDAR data with low resolution and vice versa.

The next step is the application of the k-means algorithm to the cloud trying to divide it into two clusters. It leans on the data with the added measure to build the classification. At the end of the execution, a new classification is obtained as output. The cluster with a higher mean height is the new group of classified points. Iteration provides a new class from high vegetation to low vegetation. At last, the algorithm tries to validate the results in the last step.

In step 4, the algorithm takes the results of k-means and uses Silhouette function to decide if it is a good clustering or not. Silhouette provides a measure for every point weighting the inter-clustering distance. The measure may be between -1 and 1 where value 1 is the best and -1 is the worst. In this step, a mean for all the points is calculated and if the clusters have a silhouette mean over 0.6 the clustering is considered good. Otherwise, it is considered a bad clustering and the algorithm rollbacks, changes the points classification to ground points and ends.

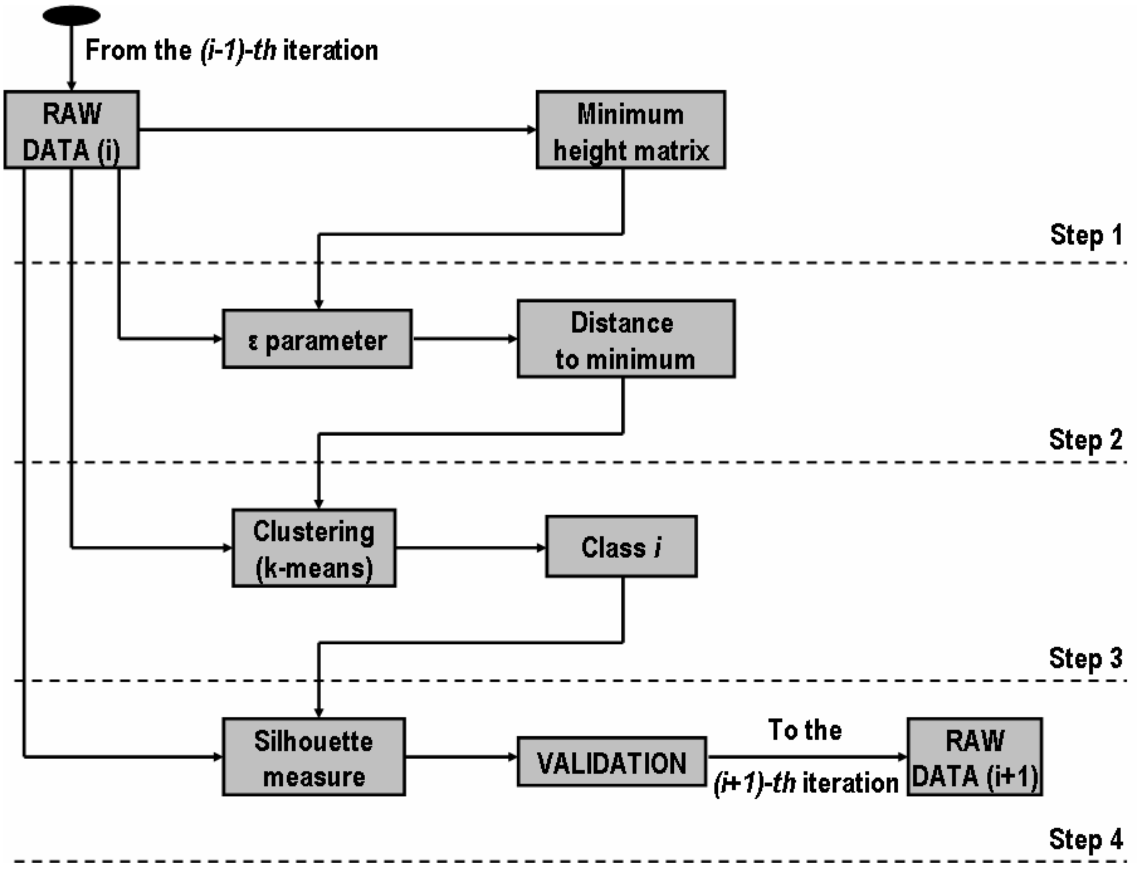

Figure 1: i-th iteration of the extracting information process.

2.2 Study area

We have chosen three study areas with different morphological (topographical) signature, with the aim of obtaining a comprehensive assessment of our algorithm, covering a forest burned and not burned area mountain range in Cerro Muriano (Córdoba) and two marshland areas, one within a Natural Setting in Isla Cristina (Huelva) and the other in Marismas del Odiel (Huelva). The task of classification is especially arduous since the terrains under study are not regular enough.

The following information was available for the study area: (i) sets of aerial photographs taken in the period from 2005 to 2007 at 1-meter resolution; (ii) medium scale lithological and land use maps in digital format; and (iii) a high-resolution (HR) digital elevation model (DEM) obtained by spatial correlation of images, including breaklines and manual edition in troubled areas. The HR DEM was used to obtain digital representations of the topographic surface of the study area, including elevation, shaded relief, and slope maps. The digital maps were exploited for visual testing, based on their morphological (topographical) appearance.

3. Results

Mediterranean woods are some of the most variable environments we can find in Europe. They usually have very little vegetation and it has a great variety of heights, forms… So the samples we have used are deemed to be difficult to filter. We have centred on vegetation because it is the most important feature for the Regional Ministry. Further studies are planning to extend the results to zones with buildings, bridges…

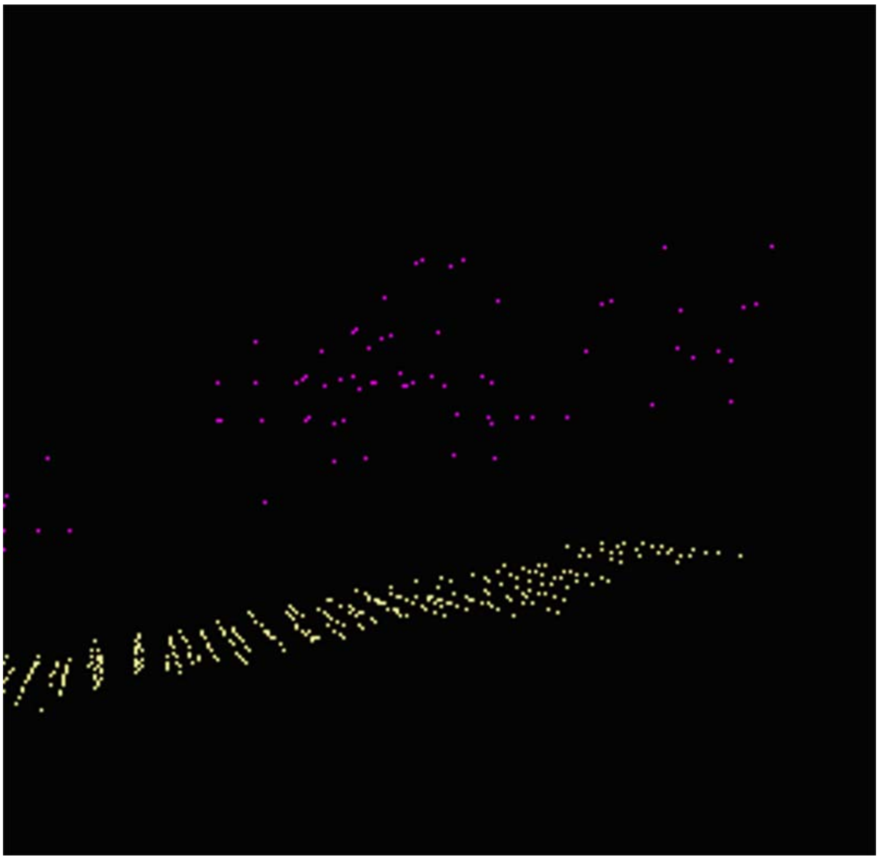

Figure 2: Part of the classification for Cerro Muriano in an early step.

The data has been processed for three zones representing the variability of the vegetation in the South of Spain. After numerous tests and estimating the ε value as a function of the inverse of the resolution of each dataset, we determined the optimum value of ε is 2 for Cerro Muriano dataset and 1 for the other two ones. The results have been compared against manually classified points. Table 1 presents the total error in every zone. The total error presents the number of misclassified points as a percentage of all studied points in the sample.

Table 2: Total error.

Zone	Number of points	Number of hits	Extractive rate (%)	Error (%)
Cerro Muriano (Córdoba)	140	112	80	20
Odiel (Huelva)	140	121	86.4	14.6
Isla Cristina (Huelva)	140	100	71.4	18.6
Total	420	333	79.28	20.72

The great error in the Cerro Muriano zone appears because it has a very low resolution. This flight was orthographic and a LIDAR sensor was mounted to make testing on the area. The low resolution affects the results even if parameter ε is got increased. In addition, the zone of Isla Cristina is an extremely difficult zone because it's a marshland. The mean height is very low and the differences between the short vegetation and the ground are hard to find. A solution to this problem is still being investigated.

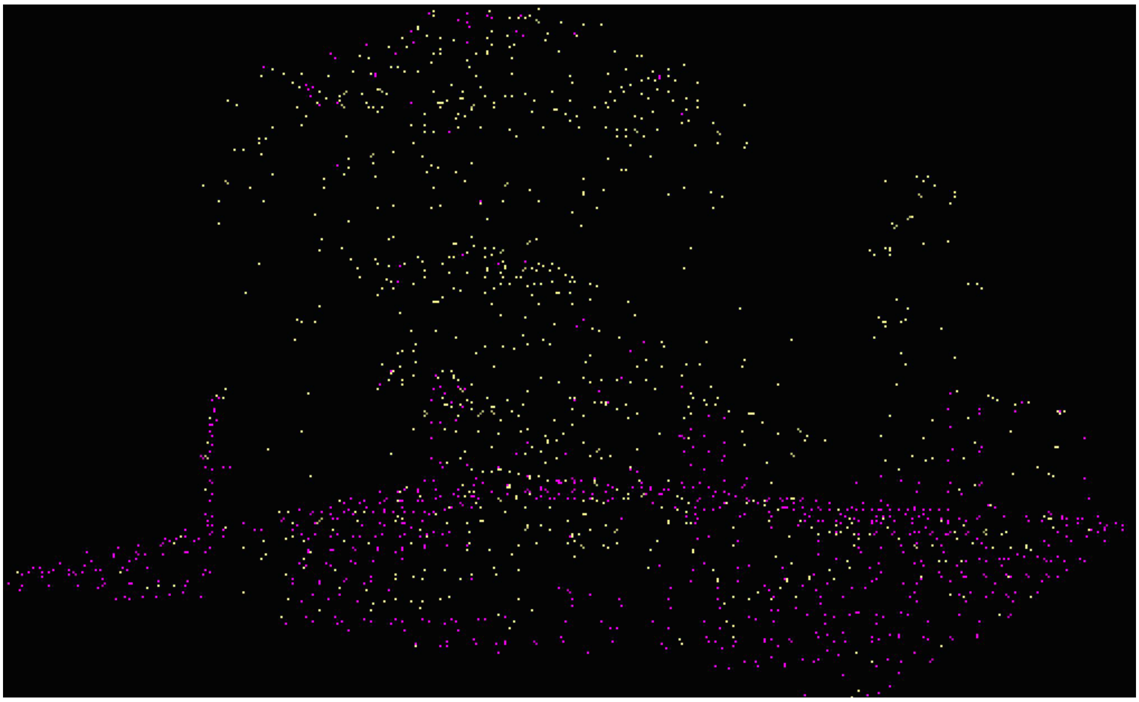

Figure 3: Example of area misclassified in Isla Cristina.

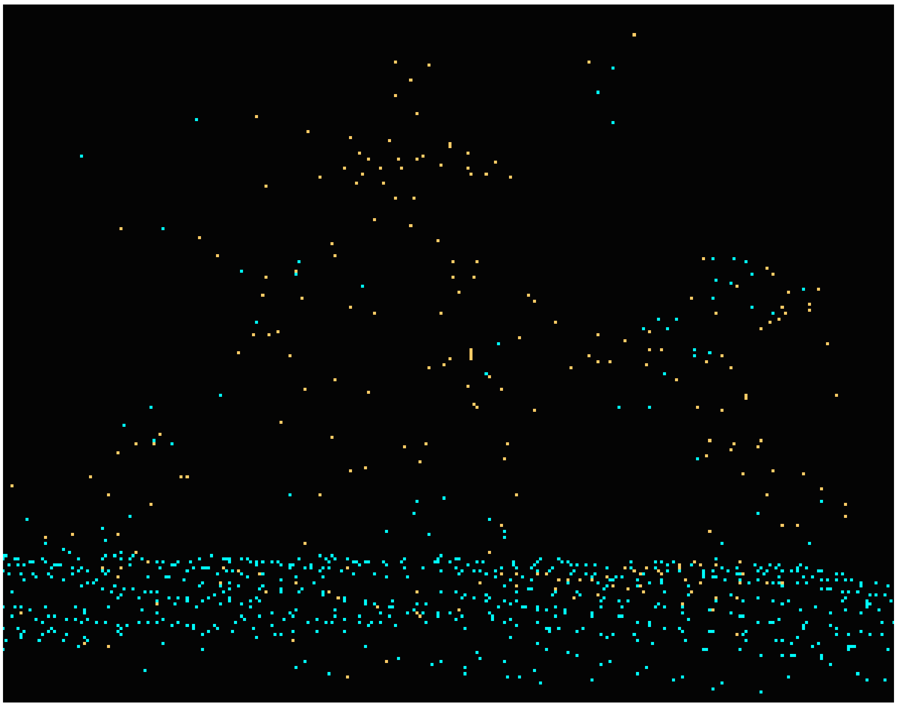

Figure 4: Part of the classification for marshland of Huelva.

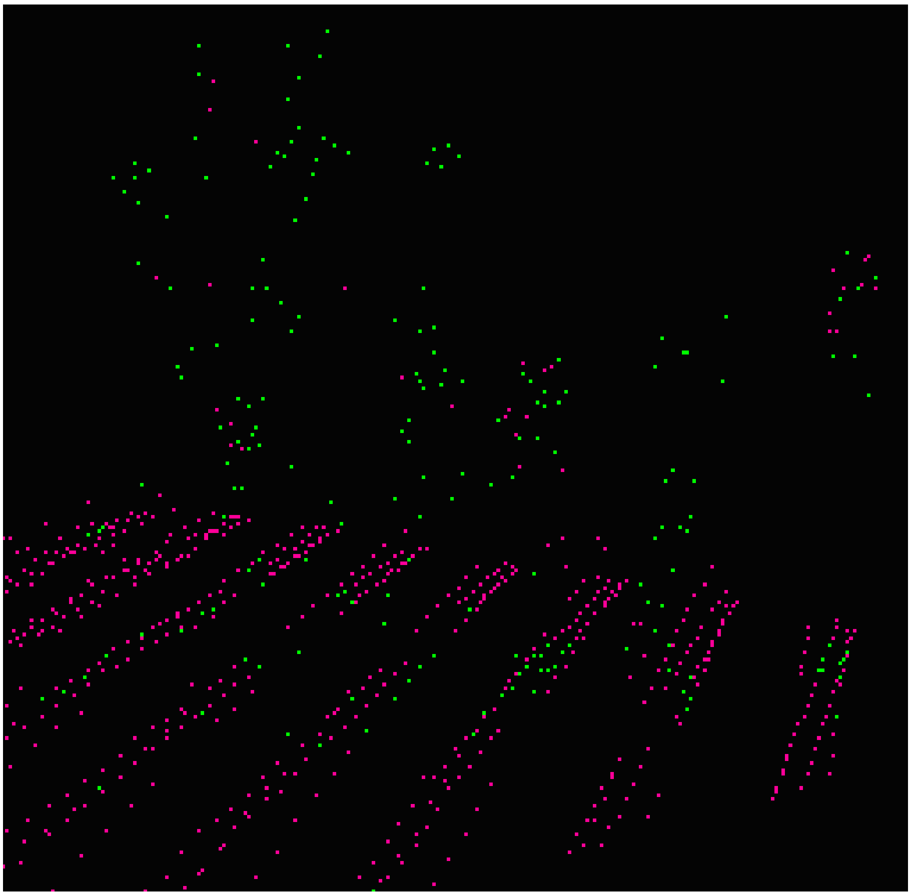

Figure 5: Example of area classified in Odiel.

One of the outstanding conclusions that can be deduced from the experiments is that our approach performs in very similar way in all landscapes. Very few algorithms are been tested on Mediterranean lands and a deep comparative study among the set of algorithms that can be found nowadays is needed to determinate if our approach can compete with them when software based on our approach is available.

3.1. Outliers

Laser scanner data sometimes contain isolated points that have large systematic error. In the developed algorithm clusters with a very small size (typically one or two points) are classified as outliers and can be deleted. If this situation appears, iteration is done again without the removed points.

4. Discussion

LIDAR data from Mediterranean woods has been analyzed in this work. To be precise, a new clustering-based approach has been proposed in order to distinguish vegetation from ground. Thus, it has been demonstrated that different kinds of profiles can be differentiated by applying a well-known data mining technique, such as k-means, integrated in a multi-step cascade process of feature extraction. A parameter is calculated in every step and subsequently used as an input of the following step. The accuracy shown is certainly promising since no extra computation, apart from the k-means, is added to the approach, achieving a low computational cost.

Concerning to future work, it can be stated that this initial division into two main classes could be very useful in order to classify miscellaneous grounds or vegetation. Moreover, this data split allows the classes to be considered and further analyzed in a different way since ground and vegetation do not show the same behaviour to laser pulses.

Acknowledgements

The authors want to acknowledge the financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084. They also want to acknowledge the support provided by the Regional Ministry for the Environment of Andalusia.

References

Bartels, M., Wei, H. and Mason, D.C., 2006. DTM generation from lidar data using skewness balancing. *In ICPR06*.

Evans, Jefrey S. and Hudak, Andrew T., 2007. A multiscale curvature algorithm for classifying discrete return lidar in forested environments. *IEEE Transactions On Geoscience And Remote Sensing*.

Fernandes, Armando M., Utkin, Andrei B., Lavrov, Alexander V. and Vilar, Rui M, 2005. Design of committee machines for classification of single-wavelength lidar signals applied to early forest fire detection. *Pattern Recognition Letters, 26:625-632.*

Filin, S., 2004. Surface classification from airborne laser scanning data. *Computers and Geosciences*.

Hamerly, G. and Elkan, C., 2003. Learning the k in k-means. *Proc. of the NIPS*.

Haugerud, R. A. and Harding, D. J., 2001. Some algorithms for virtual deforestation (vdf) of lidar topographic survey data. *International Archives Of Photogrammetry Remote Sensing And Spatial Information Sciences, 34(3):211-218.*

Morsdorf, F., Meier, E., Kötz, B. and Itten, K.I, 2004. Lidar based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management. *Remote Sensing of Environment, (92):353-362*.

Piatetski-Shapiro, G., Frawley, W. J. and Matheus, C. J., 1991. Knowledge discovery in databases: an overview. *AAAI-MIT Press*.

TerraSolid Limited, 2000. TerraScan for microStation, user's guide.

Sithole, G. and Vosselman, G., 2005. Filtering of airborne laser scanner data based on segmented point clouds**.** *ISPRS 2005*.

Vosselman, George, 2000. Slope based filtering of laser altimetry data. *IAPRS, XXXIII.*

Yan, M., and Ye, K., 2007. Determining the number of clusters using the weighted gap statistic. *Biometrics,* *63, 1031–1037*.

Zhang, K., Chen, S., Whitman, D., Shyu, M., Jianhua,Y. and Zhang, C., 2003. A progressive morphological filter for removing nonground measurements from airborne lidar data. IEEE Transactions on Geoscience and Remote Sensing, 41(4):872-882.

Link to the Regional Ministry from Enviroment of Andalusia web:
http://www.juntadeandalucia.es/medioambiente/site/web/